HUMAN ACTIVITY RECOGNITION LEVERAGING DEEP LEARNING AND FOG COMPUTING FRAMEWORKS

A SYNOPSIS

Submitted by

R. RAJA SUBRAMANIAN
(Reg. No. 201907274)

in partial fulfillment for the award of the degree of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

KALASALINGAM ACADEMY OF RESEARCH AND EDUCATION

(DEEMED TO BE UNIVERSITY)

ANAND NAGAR, KRISHNANKOIL - 626126

TAMILNADU, INDIA.

AUGUST 2023

1. Introduction

Activity Recognition (AR) uses external sensors like accelerometer or video sensors to identify different activities. It enables automatic monitoring and analysis of patient or resident activities in smart environments like smart hospitals and smart homes. AR has attracted significant interest from researchers in the fields of heath care, social care, and life care services. For instance, in a smart home, an AR system may automatically identify occupant's activities and provide activity records. These daily data can show resident's activities, which doctors/analyst might use to make further health care recommendations/ precautionary services. In today's society, ambulatory healthcare services are increasingly common, an option where pandemic illnesses affect a large number of patients. Such pandemic disease affected patients require continuous monitoring without close proximity. Especially in case of Covid affected patients, who are typically kept in quarantine for a long time, usually get depressed. It is always required to monitor/ infer the activity of such quarantined patients. People with physical ailments or musculoskeletal damages often requires such ambulatory healthcare services in terms of activity monitoring. As people with such disabilities are often unable to bounce back from some unintentional movements. An AR system can identify falls or abnormal activity patterns in older persons, in particular, and notify or educate caretakers. Activity recognition is one of the important criteria in smart city applications from a surveillance standpoint, along with ambient health services.

Techniques for activity feature extraction, modelling, and recognition are part of the fundamental approach for activity recognition. A difficult challenge is video-based AR [1], since it must take into account a human's entire range of motion and does not follow to strict syntax like hand gestures or sign languages. Therefore, to accurately define motions in this regard, a thorough model of the entire human body is required. Due to its useful uses, several researchers have been investigating video-based AR systems. According to motion characteristics, video-based AR may generally be classified into two categories, viz. marker-based and vision-based. The earlier model [2] is built on a wearable optical marker-based Motion Capture (MoCap) device, which is popular because it has the benefit of correctly recording complicated motions. The optical sensors must be attached to the body, which has the issue of requiring several camera

setups. The other is marker-free and based on depth video cameras. The lack of wearable tracking markers is drawing greater attention to this method these days, making it simple to implement the AR system in practical applications.

As for the classification methods, HMMs (Hidden Markov Models) have till now been extensively employed [3] in several AR systems since they can decode temporal patterns. However, due to its capacity to understand complex patterns, deep learning for recognition is currently receiving a lot of attention. Deep learning, in its basic form, refers to neural networks that use non-linear data processing layers for feature categorization. These layers, which operate in a hierarchical structure, handle the results of the earlier levels. Many conventional approaches in computer vision have been improved by deep learning techniques. Implementing a simple human action and a particular transition action are both components of complex human activities. A nonlinear time series classification with labelled data is modelled leveraging conventional deep learning methods like SVM (Support Vector Machines) [4], Random Forest, etc. and non-conventional deep learning methods like ANN (Artificial Neural Network), CNN (Convolutional Neural Network) [5], RNN (Recurrent Neural Network) [6], among others. A number of earlier works have investigated the problem of activity recognition. Traditional methods provide a time-consuming difficulty since they need extensive feature engineering and human object detection/ extraction. On the other hand, deep learning approaches can automatically learn the characteristics from the data and are therefore more suitable for the task of identifying complicated human objects and behaviours.

In two aspects, deep learning approaches are particularly promising for solving AR needs. First, compared to standard recognition approaches, performance can be greatly enhanced. Second, deep learning techniques may be used to find characteristics related to the dynamics of human motion, ranging from straightforward motion encoding in lower levels to more complicated motion dynamics in higher layers of the network.

2. Literature Survey

There are two methods for carrying out this activity recognition task: body area sensors and vision cameras. Sensors are typically attached to the patient's body to track their movements, accelerations, and other physiological characteristics. To recognise

human activity, cameras use image processing and learning frameworks. The task of online activity recognition has been feasible with the development of numerous Internet of Medical Things (IoMT) and learning technologies. There are two ways to infer activities: unimodal or multimodal. Based on the types of sensors used for activity recognition, a classification is made. The unimodal model uses actions, rules, statistics, and spatial/ temporal data generated from image sensors as a single source as stated by Islam et al. [4]. The later way of activity recognition gets new elements for activity recognition from the subject's interactions and behaviours. The cameras can act as image sensors (UAV or CCTV cameras) placed throughout the environment. The difficulties in recognising activity from image sensors can be divided into two categories: efficiency and effectiveness. Effectiveness is concerned with the model's precision in predicting the suitable human activity by applying learning algorithms to the collected image frames. Activity Inference in medical applications are typically time-critical. After the deadline, the accurate result is no longer helpful. Hence efficiency is a crucial performance factor in activity inference frameworks.

Typical applications of the kind are usually run-in cloud computing architectures with the model being constructed and loaded in the cloud. The captured live video frames during scoring are sent to cloud for analysis and subsequent action inference. The result from the cloud is then projected to the edge device and subsequent app. This type of traditional architecture is subjected to a huge data transit time. The transfer of huge video frames from edge to cloud and the projection of results from cloud to edge often affects the overall response time of the model. In order to reduce the response time, alternate ways to store the model is required.

According to Tuli et al. [7] Fog Computing is visualized as one of the prominent sub-domains of Internet of Things. It projects to be a vital resolver of the issues faced by cloud computing applications in terms of response time. Fog layer is a simplified version of cloud layer. The fog layer composes of fog nodes orchestrated in groups of varied sizes depending on the sub task assigned with respect to the application in question. The fog/ gateway layer groups the fog nodes into clusters with a cluster head or master governing several slave fog nodes. The fog master with its local fog nodes is visualized as simple cloud. These fog nodes, unlike cloud machines, are typically

commodity systems with minimal storage and computing capabilities, just enough to the sub-task assigned. The fog master of the cluster is responsible for the scheduling, load-balancing and maintenance of its nodes. The reliability of the orchestration is ensured by a secondary fog master, which assumes action with any disruptions of the primary head.

Various algorithms exist in the literature for inferring commonality among image frames including structural descriptors, textural descriptors, among others. Wang et al. [3] experimented the Local Binary Patterns, Local Ternary Patterns, Local Phase Quantization that leverage on binning operations to generate descriptors. The resultant histogram-based descriptors are analysed for finding discriminations among image samples. These feature descriptors suffer from computational complexity as the image space increases. The activity recognition tasks have a considerably large input space, depending upon the number of activities. Hence these kind of binning techniques cannot be employed effectively for the underlying time critical task, which finds its application in medical fields. Image transforms such as discrete cosine transform, randon transform, Arnold transform also serves as feature descriptors. Unlike textural feature descriptors, the transform based featured do not leverage on binning techniques. On the contrary, these transforms are subjected to the defects caused by noise and outliers. As the images captured for activity recognition is subjected to noise and the output of object detection algorithms like YOLO described by Zhang et al. [8], the human frames, can include such defects. Hence the accuracy levels on these transform features are not par enough for activity recognition tasks. Various deep learning frameworks including CNN models with pretrained ResNet of Wang et al. [1], AlexNet of AlDahoul et al. [9], GoogleNet, LieNet of Zhu et al. [10] models are proposed for image recognition. The common classification techniques used prevalently in the literature are SVM (Islam et al. [4]), Linear classifier (Mliki et al. [11]), maximum likelihood estimation (Simonyan and Zisserman [12]), among others. There requires a significant algorithm that provides a high degree of accuracy in comparing image frames at a faster rate. In order to reduce the number of comparisons, change detection between consecutive frames can be done to predict the next activity or retain the previous activity. The algorithm can also be aided with the fog computing framework to aid in efficiency factor.

2.1. Datasets for Activity Recognition

Hollywood2 Action Dataset

Hollywood2 action dataset, described by Marszalek et al. [13], is a dataset captured from fixed camera. Thus, these video sequences do not require scene stabilization, to counter the camera movements. The dataset composes 1707 video sequences from around 69 movies. The video sequences evidence 12 different activity classes including answer_phone, handshake, drive_car, eat, get_out_of_car, fight_person, hug_person, kiss, run, sit_down, sit_up, and stand_up. The video dataset occupies storage of around 15 GB playing for 20.1 hours.

KTH Dataset

KTH dataset, described by Schuldt et al. [14], also composes video sequences captured from static cameras. This dataset, commonly used for activity recognition research, has 2391 video sequences captured from 25 fps camera in same background. The video is contributed by 25 people with 6 activities namely boxing, clapping, hand-waving, jogging, running, walking.

UCF-ARG Dataset

UCF-ARG dataset, described by Nagendran et al. [15], is a benchmark dataset for conducting experiments on human detection and corresponding activity recognition from video sequences captured from aerial cameras. The dataset possesses videos contributed by twelve people witnessing 10 human activities. The dataset is multifaceted with its constraints including the extreme ego-motion, the lighting levels and the camera height variation. In our experimentation, we focus on the recognition of following five activity classes: throwing, running, digging, waving and walking.

2.2. Challenges in Activity Recognition

1. The recognition task should be carried out without the knowledge of the stakeholders being monitored, in which case, body sensors cannot be used
2. Camera based activity recognition usually takes a huge computational complexity, hence response time need to be controlled, with respect to the application in question
3. Time-critical applications require faster recognition of activities

4. Activity recognition schemes leveraging cameras can be operated in a closed/ confined environment only and cannot be applied in open spaces, where moving cameras/ drones are required
5. Moving cameras/ drone captured videos are usually liable to scene stabilization issues, hence appropriate pre-processing should be carried out before running the inference model
6. Feature extraction algorithms typically consume huge time and storage to generate descriptors for various activities pertaining to the videos frames under study
7. Loading a naive learning model trained on the video data set is expensive and cannot be reliable in applications that requires continuous monitoring

3. Objectives of the Research

The primary objectives of the work are:

(i) To generate an efficient learning model architecture for recognizing human activity from physiological sensors

(ii) To develop a activity recognition framework using fixed cameras in confined environment

(iii) To develop a accelerated deep learning framework for recognizing activity of human beings using UAVs and fixed cameras

(iv) To analyse the performance of time-critical activity recognition process in fog environment

4. Research Problem

Ambulatory healthcare services have become prevalent in today's world where many patients suffer from pandemic diseases. Physically or mentally challenged people, childhood and elderly people require continuous monitoring. It is always safer to be proactive in such people's health. During isolation, these categories of people have higher health risks with respect to their activities. Movements or actions of mentally challenged people can jeopardize their health, as they cannot control themselves. Similarly, children, aged people and physically challenged people cannot recover themselves from some accidental movements. In addition, people suffering from pandemic diseases require to have a self-quarantine. Such isolation has various mental effect on the patients. Hence it is required to monitor and safeguard the patients. Hence it is required to monitor and safeguard the patients. Due to lack of physicians or care takers in the

remote environments, their require an ambient system which can automatically recognize human activities and send appropriate notifications to required sources in the case of vulnerabilities. Such activity monitoring task can be performed in two ways, viz. the body area sensors and the vision cameras. Sensors usually get affixed in body parts of the patient and monitor movements, accelerations and other physiological parameters. Cameras leverage the image processing and learning paradigms to recognize activities of human beings. With the development of various IoMT and learning technologies, the task of online activity recognition become practical. In this research, we propose four different frameworks for Human Activity Recognition, viz. HARFog, HARDeep, HARGen and HARNet.

4.1 HARFog: Activity Recognition Framework for Physiological Sensor Inputs

We propose a Cloud-fog interoperable model, named as HARfog, composing the edge sensors acquiring data from human, the set of gateway nodes with deep learning models and the central cloud.

The mHealth dataset [16] consists 23 attributes of sensor information classified into 12 activities: standing, sitting, lying down, walk, climbing stairs, waist bends forward, frontal arm elevation, knee bending, cycling, jogging, running, jump front & back. The data present in each node uses the knowledge base acquired during training. The predictive analytics is performed parallelly by multiple nodes. The bagger combines the results among the nodes through maximum voting scheme. The training and testing accuracies of the prediction model, with respect to the number of edge nodes is depicted in figure 1a and figure 1b respectively. The accuracy is calculated as the percentage of total activities correctly classified. It is evident that the training time decreases as the number of nodes increases. The distributed training knowledge across the edge nodes decreases the testing accuracy. The timing parameters, latency and execution time, arbitration time and jitter, for different fog scenarios is observed, by running the model in fogbus framework. Five scenarios including Master only, 1 edge node, 2 edge node, ensemble mode and cloud are observed.

Other works that proposed the architectures for HAR [17, 18, 19] leverage pri-

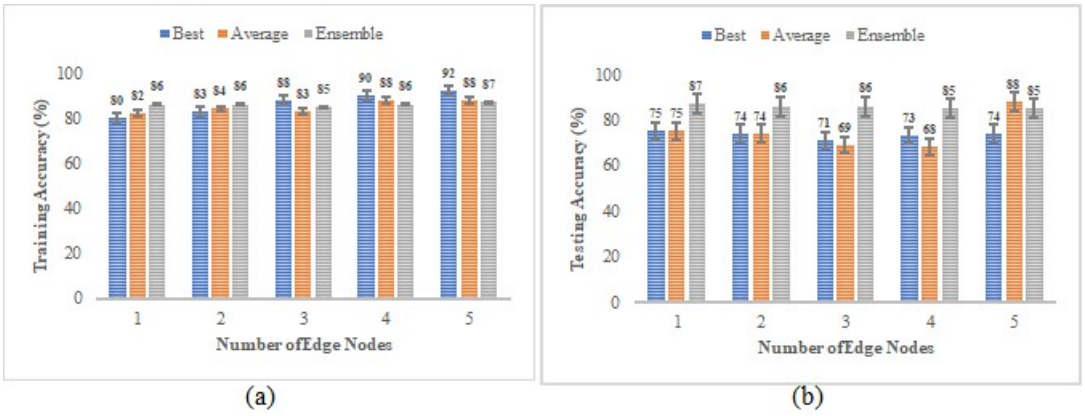

Figure 1: (a) Training Accuracy, (b) Testing Accuracy, with number of edge nodes

marily on learning paradigms confined to the data under study. The evaluation of the HAR model with respect to timing constraints is scarce, to the best of our knowledge. HARfog is able to perform analytics on the data for time critical tasks, as it leverages fog-assisted cloud computing framework. With ensemble algorithm and multiple edge nodes support, accuracy of the model is higher than the existing state-of-the-art HAR models.

4.2 HARDeep: Deep Learning Framework for HAR from Video Inputs

HARDeep, consists of three modules: (i) An optional Scene Stabilization module (ii) Identification of useful frames (iii) Inference of human activity. Scene Stabilization is carried out to distinguish the motions caused by object displacements and the motions caused by camera displacements. The latter motions require stabilization to counter the motions caused by camera movements. If there exists a motion between two consecutive frames, an image pixel in initial frame (I_{K_t}) and that in the consecutive frame $(I_{K_{t+dt}})$, will have a displacement during the time interval dt. The human detection task, being analogous to object detection in the video sequences, is carried out using You Only See Once (YOLOv3) network model [20, 21].

An Ensemble model combines the efforts of multiple learning models and accumulates the results by weighting their decisions. Thus, ensemble algorithm will have better performance than the atomic ones. In order to build a deep ensemble model, the performance of five pretrained models: Alexnet, Vgg-16, Resnet-50, Resnet-101 and Googlenet are evaluated and the best three models: Resnet-50, Resnet-101, Googlenet

are collaborated to develop an ensemble classifier leveraging majority voting scheme. These imagenet models have a common capability of describing an image with 1000 features in one of the layers. The layer from which the features are extracted are as follows: Loss3 in Googlenet, FC-1000 in Resnet-50 and Resnet-101.

The inference accuracy of the HARDeep framework is evaluated against various standard and complex video datasets including Hollywood 2 dataset, KTH dataset and the UCF-ARG dataset, containing videos captured from UAV. Table 1 depicts the evaluation results of HARDeep on the three activity datasets under study. It exhibits a recognition accuracy of 97.13%. The proposed framework recognizes activities from KTH dataset at the accuracy level of 96.75%. The framework is also experimented against dynamic aerial camera captured video sequences from UCF-ARG dataset, and an inference accuracy of 80.72% is obtained. It shows that the proposed model exhibits better recognition accuracy than the existing models.

Table 1: Evaluation results of HARDeep on three activity datasets under study

Datasets	TP	TN	FN	FP	Precision	Recall	Accuracy
Hollywood2	3282	83588	1406	1179	0.7357	0.7001	0.9713
KTH	18939	79198	1313	1995	0.9047	0.9352	0.9675
UCF-ARG	14895	57033	9035	8087	0.6481	0.6224	0.8072

The proposed model, HARDeep, for activity recognition gives better accuracy with three standard video datasets, viz. Hollywood2, KTH and UCF-ARG. Since activity recognition models finds use cases in time-critical applications, the model complexity in terms of response time need to be estimated. We execute the ensemble HAR model in fogbus [7] framework. Fogbus, being the state-of-the-art framework for modelling fog-based cloud computing architectures, is leveraged. The efficiency of the proposed HAR model is estimated in terms of latency, execution time, jitter and arbitration rate, for four different fog scenarios

4.3 HARGen: Deep Genetic Algorithm for HAR

HARGen framework orchestrates the Activity inference task in three folds, viz. Object detection, Change detection and Activity Inference. The complex activity inference model used for recognizing the activity is called, only when there is a significant activity change detected in the consecutive video frames. Detecting changes in consecutive

frames, requires comparing only two video frames, whilst the activity inference model requires comparison of the image frames under study with all possible activities in the data set and infer the correct activity. Hence the complex inference tasks are avoided for frames which does not have significant change in its pixels.

The initial sequence of videos frames requires a scene stabilization phase [22], if they are captured through UAV cameras. This discriminates the motion corresponding to the objects and the cameras in the video sequence. The frames obtained from the video sequence is subjected to an object recognition algorithm: YOLO. The detected human objects will be inputted to the change detection algorithm. The change detection phase is implemented over a accelerated genetic algorithm. The proposed change detection algorithm is orchestrated with three steps: (i) Denoising and difference image generation through BM3D, (ii) Initial classification of image frames leveraging FCM and (iii) Accelerated Genetic Algorithm employment for change map generation. The changed/ new frame will be subjected to the activity recognition algorithm. The unchanged frames will be labelled with the same activity of its predecessor. Thus, reducing the number of frames being analysed for activities. Object detection, change detection and activity recognition are the core components of the proposed architecture. The ultimate aim is to infer activity at a rapid rate, as the application being classified as time-critical. Hence the algorithms are chosen and orchestrated to run efficiently. The performance of HARGen is compared over state-of-the-art activity recognition schemes as depicted in Table 2.

The model is executed in fogbus [7], the framework for modeling fog-assisted cloud models. Timing parameters including latency, execution time, arbitration time and jitter are measured under four fog scenarios: Master-only, 1 Edge node, 2 Edge nodes and Cloud.

4.4 HARNet: Yoga Posture Recognition Framework

The proposed idea of the research work is to create an interactive toy leveraging the state of art technologies. The toy aims to improve mental ability, social interactivity, and adaptive to the environment. Learning Yoga is incorporated with the toy to make the kids learn and practice yoga. The Yoga detection model is inserted in the toy to detect the yoga pose performed by the kid and in return instruct the kid with the mistake

Table 2: Comparison of the recognition accuracy of proposed technique with the state-of-the-art activity recognition technique

Techniques	Hollywood2	KTH	UCF-ARG
Wang et al. [26]	64.30	-	-
Lan et al. [27]	68.00	-	-
Jiang et al. [28]	-	95.77	-
Tian et al. [29]	-	94.50	-
Ryoo et al. [30]	-	91.10	-
Zhang et al. [31]	-	94.00	-
Bregonzio et al. [32]	-	93.17	-
Thi et al. [33]	-	94.67	-
Seo et al. [34]	-	95.1	-
Mehrsan et al. [35]	-	95.00	-
Mikolajczyk et al. [36]	-	95.30	-
Ke et al. [37]	64.60	-	-
Jain et al. [38]	-	62.96	-
Islam et al. [4]	87.00	75.00	-
Burghouts et al. [39]	-	-	57.00
Mliki et al. [11]	-	-	68.00
Proposed Deep Genetic Algorithm	98.42	97.83	81.40

to perform in the right manner. The Yoga detection model requires action recognition framework and we leverage our HAR assisted model to incorporate suitable deep learning techniques i.e., ResNet50 which is a transfer learning model to train the toy with the required yoga asanas. The Yoga-82 [40], an open-source dataset is used to train the model. In this research, we consider 3 different yoga asanas namely Bitilasana, Chaturanga Dandasana and Ustrasana. The camera incorporated in the toy captures the images and sends them to the model to get the classified image as the respective yoga posture. Further, if any corrections to be done, the detection model identifies the errors and instructs the kid as required.

The implementation of the project can be classified into three categories. (i) Implementation of HARNet (Software module) (ii) SensInt Implementation (Hardware Module) (iii) Integration of HARNet and SensInt implementation. The Implementation of the HARNet explains clearly the dataset taken to train the yoga asanas and the modeling of the yoga gesture detection model. The SensInt implementation clearly explains the working principle of the sensors and other electrical components used to make the

physical toy. The integration of hardware and software explains the consolidating part of both hardware and software parts to prepare the whole toy and the additional components added to the proposed toy. Also, the integration part discusses the IoT architecture used.

The Deep Learning module plays a crucial role in the implementation of the proposed toy. The toy consists of learning and instructing Yoga which needs a Yoga posture detection model for analysing the yoga done by the kid as a captured image. Then, the yoga asana which is performed needs to be identified. Initially, we consider the Yoga-82 Dataset [40] which is an open-source dataset available for classifying the human poses. The dataset generally consists of 82 different yoga poses or classes. Each class consists of images in the range of 40 to 90 images. The images in each class are taken in a different direction, and angle, and posed in different steps with various backgrounds. This could help us in detecting the real-time as the kids may do their yoga practice in their home environment.

Hence, the dataset suits well for the defined research area. From the whole set of data, we consider only 3 different classes of yoga asana which are required for incorporation in the toy, and yoga asanas can create peace in the mind and good exercise to improve their mental as well as physical health. The yoga asanas considered are namely Bitilasana, Chaturanga Dandasana and Ustrasana. The 3 different yoga asana classes are manually separated from the whole data and used for training the model. Each class consists of 70 to 80 images in different directions, angles, and in the pose of making the yoga (steps to be directed) and also in various backgrounds. The model is tested with the validation image and test images to get good results. The training accuracy of the model ranges from 88.38% to 99.64%. The test accuracy received by the model ranges from 93.75% to 98.52% of accuracy giving the loss values of range 15.73% to 0.305%. The accuracy of the model increases with the increase in epochs.

Ultimately, the motive of the approach is to promote and enhance the learning of autistic kids through the medium of toys. And hence, we assimilate the state of art technologies into the toy, which turns out to be effective and imperative as soon the toy gets embedded with the latest technology. The learning through this toy is creative, visually appealing, playable, talkable, and sensible. These are the essential abilities required to

make the autistic engageable and attracted to the teaching-learning method through the toys and play models. Therefore, the toy satisfies the needs including the latest trends in technology. Further, the project is to be improved with the supplementary resource embedding to have an improved version of hardware and software modules including the integration part of the proposed toy. The efficient results obtained through hardware and software modules increase the advantages and effectiveness of the toy contributed to the community. Even though the solution has ethical and sustainable constraints, we try to improve our toy in further research. The idea of the future work is to enhance the application interface using the IoT Architecture of edge computing related to the resources.

The research has many open problems which could be addressed in the future. Activity recognition or gesture recognition will have wide applications in surveillance and smart city requirements. In autonomous vehicles, path planning aided through activity recognition can be performed in crowded areas for safety requirements. In addition to using cameras, RADAR and LIDAR based object tracking and subsequent recognition of activities can be performed. Such RADAR based/ LIDAR based activity recognition overcomes the privacy invasive camera based inputs and the hard wearable physiological sensors.

5. Organization of the Thesis

Chapter 1, “Introduction”, presents the various applications of Human Activity Recognition (HAR) models, the motivation for pursuing research in HAR, the underlying challenges, objectives and contributions of our research.

Chapter 2, “Literature Survey”, depicts the taxonomy of various image recognition models, the overview of fog computing frameworks, the datasets that are leveraged in the research for modelling HAR frameworks.

Chapter 3, “HARFog: HARfog: Activity Recognition Framework for Physiological Sensor Inputs”, portrays the design of HARFog framework for recognizing human activities from physiological sensors and evaluation of the framework on fogbus.

Chapter 4, “HARDeep: Deep Learning Framework for HAR from Video Inputs”,

explains the design of HARDeep framework that learns the video frames from cameras and UAVs to recognize human activities.

Chapter 5, “HARGen: Deep Genetic Algorithm for HAR”, discusses the accelerated learning framework that combines genetic and deep learning paradigms for pursuing HAR.

Chapter 6, “HARNet: Yoga Posture Recognition Framework”, elaborates the implementation of HARGen model on a Yoga activity recognition task, as an application teach Yoga for Autistic kids.

Chapter 7, “Conclusion and Future Scope”, summarizes the overall contributions of the research and scope for future advancements.

6. References

[1] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang and P. Ogunbona, ConvNets-based action recognition from depth maps through virtual cameras and Pseudocoloring, in: *Proceedings of the 23rd ACM International Conference on Multimedia*, (2015), 1119–1122.

[2] J. Yue-Hei, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 4694–4702.

[3] H. Wang and C. Schmid, Action recognition with improved trajectories, in: *IEEE International Conference on Computer Vision*, (2013), 3551–3558.

[4] N. Islam, Y. Faheem, I. Ud Din, M. Talha, M. Guizani and M. Khalil, A blockchain-based fog computing framework for activity recognition as an application to e-Healthcare services, *Future Generation Computer Systems*, **100** (2019), 569-578.

[5] Y. Li, W. Li, V. Mahadevan and N. Vasconcelos, Vlad3: encoding dynamics of deep features for action recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, (2016), 1951–1960.

[6] Y. Shi, Y. Tian, Y. Wang, and T. Huang, Sequential deep trajectory descriptor for action recognition with three-stream CNN, *IEEE Transactions on Multimedia*, **19**(7) (2017), 1510–1520.

[7] S. Tuli, M. Redowan, T. Shikhar and B. Rajkumar, FogBus: A Blockchain-based Lightweight Framework for Edge and Fog Computing, *Journal of Systems and Software*, **154** (2019), 22-36.

[8] Y. Zhang, Y. Shen and J. Zhang, An improved tiny-yolov3 pedestrian detection algorithm, *Optik*, **183** (2019), 17-23.

[9] N. AlDahoul, M. Sabri, A. Qalid and A. Mansoor, Real-time human detection for aerial captured video sequences via deep models, *Computational intelligence and neuroscience*, (2018).

[10] W. Zhu, C. Lan and J. Xing, Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep Lstm Networks, *arXiv Prepr, AAAI, Phoenix, Arizona, USA*, **2** (2016), 8-15.

[11] H. Mliki, F. Bouhlel and M. Hammami, Human activity recognition from UAV-captured video sequences, *Pattern Recognition*, **100**(1) (2019), 107-140.

[12] K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, (2014).

[13] M. Marszalek, I. Laptev and C. Schmid, Actions in context, in: *IEEE Conference on Computer Vision and Pattern Recognition*, (2009).

[14] C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: a local SVM approach", in: *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004*, **3**, (2014) 32–36.

[15] A. Nagendran, D. Harper and M. Shah, UCF-ARG dataset, University of central florida, `http://crcv.ucf.edu/data/UCF-ARG.php` (2010) (Accessed: 2023-02-16).

[16] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. Holgado, S. Lee, H. Pomares and I. Rojas, Design, implementation and validation of a novel open framework for agile development of mobile health applications. *Biomedical Engineering Online*, **14**(S2:S6) (2015), 1-20.

[17] L. Cao, Y. Wang, B. Zhang, Q. Jin and A. Vasilakos, GCHAR: An efficient Group-based Context–aware human activity recognition on smartphone. *Journal of Parallel and Distributed Computing* (2017), 181-196.

[18] F. Ordonez and D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, **16**(115) (2016), 126-137.

[19] P. Mamoshina, A. Vieira, E. Putin and A. Zhavoronkov, Applications of deep learning in biomedicine. *Molecular pharmaceutics*, **13** (2016), 1445–1454.

[20] S. Lu, B. Wang, H. Wang, L. Chen, M. Linjian and X. Zhang, A real-time object detection algorithm for video, *Computers and Electrical Engineering*, **77** (2019), 398-408.

[21] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, (2018).

[22] J. Barron, D. Fleet and S. Beauchemin, Performance of optical flow techniques, *International journal of computer vision*, **12**(1) (1994), 43–77, .

[23] J. Uijlings, K. Sande, T. Gevers and A. Smeulders, Selective search for object recognition, *International journal of computer vision*, **104**(2) (2013), 154–171, .

[24] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 770–778.

[25] N. Chen, Y. Chen, S. Song, C. Huang and X. Ye, Smart urban surveillance using fog computing, in: *IEEE/ACM Symposium on Edge Computing (SEC)*, (2016), 95–96.

[26] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin and J. Wu, Human action recognition by learning spatiotemporal features with deep neural networks, *IEEE Access*, **6** (2018), 17913–17922.

[27] Z. Lan, M. Lin, X. Li, A. Hauptmann, and B. Raj, Beyond gaussian pyramid: multi-skip feature stacking for action recognition, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA*, (2015), 204–212.

[28] Z. Jiang, L. Zhe and L.S. Davis, Recognizing human actions by learning and matching shape–motion prototype trees, *IEEE Trans. Pattern Anal. Mach. Intell.* **34** (2012), 533–547.

[29] Y. Tian, L. Cao, Z. Liu and Z. Zhang, Hierarchical filtered motion for action recognition in crowded videos, *IEEE Trans. Syst. Man Cybern.* **42**(3) (2012), 313–323.

[30] M.S. Ryoo and J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, *IEEE International Conference on Computer Vision (ICCV)*, (2009), 1593–1600.

[31] Y. Zhang, X. Liu, M.C. Chang, W. Ge and T. Chen, Spatio-temporal phrases for activity recognition computer vision, in: *European Conference on Computer Vision (ECCV), Springer*, (2012), 707–721.

[32] M. Bregonzio, G. Shaogang and X. Tao, Recognising action as clouds of space–time interest points, *Computer Vision and Pattern Recognition (CVPR)*, (2009), 1948–1955.

[33] T.H. Thi, L. Cheng, J. Zhang, L. Wang and S. Satoh, Integrating local action elements for action analysis, *Computer Vision Image Understanding* **116**(3) (2012), 378–395.

[34] H. Seo and P. Milanfar, Action recognition from one example, *IEEE Trans. Pattern Anal. Mach. Intell.*, **33** (2011), 867–882.

[35] M. Roshtkhari and M. Levine, Human activity recognition in videos using a single example, *Image and Vision Computing*, **31**(1) (2013), 864-876.

[36] K. Mikolajczyk and H. Uemura, Action recognition with appearance–motion features and fast search trees, *Computer Vision Image Understanding* **115**(3) (2011), 426–438.

[37] Y. Ke, R. Sukthankar and M. Hebert, Efficient visual event detection using volumetric features, in: *Tenth IEEE International Conference on Computer Vision, ICCV'05*, **1** (2005), 166–173.

[38] M. Jain, H. Jégou and P. Bouthemy, Improved motion description for action classification, *Front. ICT* (2016).

[39] G. Burghouts, A. van Eekeren and J. Dijk, Focus-of-attention for human activity recognition from uavs, in: *Electro-Optical and Infrared Systems: Technology and Applications XI*, *International Society for Optics and Photonics*, **9249** (2014), 924-941.

[40] M. Verma, S. Kumawat, Y. Nakashima and S. Raman, Yoga-82: A New Dataset for Fine-grained Classification of Human Poses, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2020), 4472-4479.

7. List of Publications

1. R. Raja Subramanian, V. Vasudevan, "A deep genetic algorithm for human activity recognition leveraging fog computing frameworks", *Journal of Visual Communication and Image Representation*, **77** (2021), 103–132.

2. R. Raja Subramanian, V. Vasudevan, "HARDeep: design and evaluation of a deep ensemble model for human activity recognition", *Int. J. Innovative Computing and Applications*, **14**(3) (2023), 155–166.

3. Raja Subramanian, R., Vasudevan, V., "HARfog: An Ensemble Deep Learning Model for Activity Recognition Leveraging IoT and Fog Architectures" *In: Gunjan, V.K., Zurada, J.M. (eds) Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough. Studies in Computational Intelligence, Springer, Cham*, **956** (2021), 127-136.